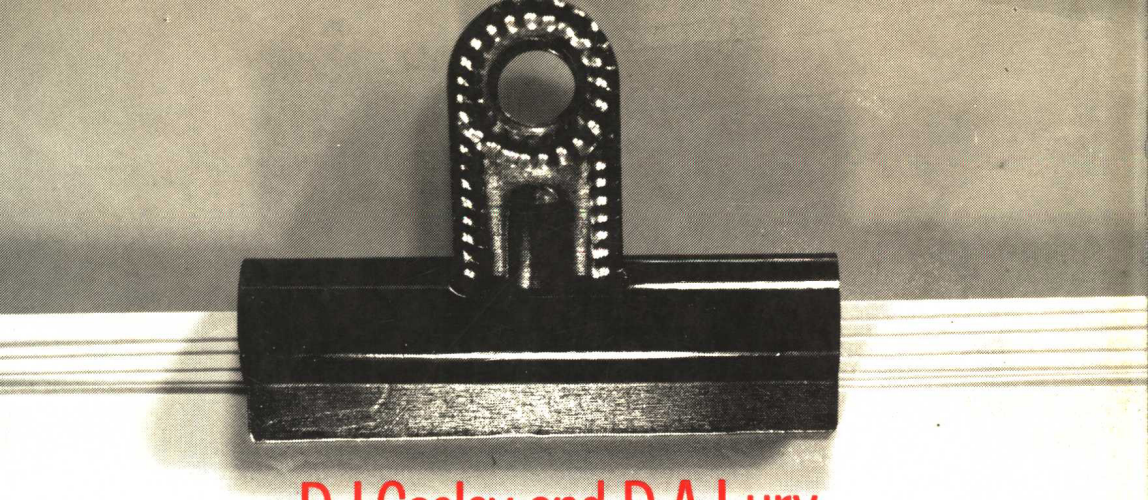

D.J.Casley and D.A.Lury

Data Collection in Developing Countries

Data Collection in Developing Countries

D.J. CASLEY
AND
D.A. LURY

CLARENDON PRESS OXFORD

Oxford University Press, Walton Street, Oxford OX2 6DP
London Glasgow New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar Es Salaam Cape Town
Melbourne Auckland
and associate companies in
Beirut Berlin Ibadan Mexico City

Published in the United States by
Oxford University Press, New York

First published 1981
First published in paperback 1982

British Library Cataloguing in Publication Data
Casley, D J
Data collection in developing countries
1. Sampling (Statistics)
2. Underveloped areas—Social surveys
I. Title
309'.07'23 HA31.2 80-41054

ISBN 0-19-877123-1
ISBN 0-19-877124-X Pbk

Printed in the United States of America

DATA COLLECTION
IN
DEVELOPING COUNTRIES

To
Jill
and
Toni

PREFACE

Our purpose in writing this book is set out in the Introduction. We decided to write it in Nairobi in 1978 when we renewed a working relationship that had started in Uganda twenty years earlier. Looking back, and reviewing the current situation, we thought that a book concentrating on practical problems of data collection in developing countries would be helpful. We know, of course, that many of the problems are problems without solutions; but continuous effort directed by common sense and experience can do much to reduce difficulties. We have set out what we have learned, in the hope that this will help others avoid our mistakes.

The quotations at the beginnings of the chapters are taken from the cases of Sherlock Holmes, reported by A. Conan Doyle.

Chapters have been read and commented on by our colleagues, A. Bebbington, J. Dobby, M. F. Fuller, M. A. Katouzian, E. Oxborrow, S. C. Pearce, A. Rutherford, and P. Stirling of the University of Kent; and T. Marchant, R. D. Narain and S. Narula of the FAO. We are grateful for their assistance; but they are not, of course, responsible for our errors.

One of us (D. Casley) wishes to thank the FAO of the United Nations for permission to publish this book whilst a staff member, and to acknowledge the use of notes of lectures given on behalf of the FAO. We emphasize, however, that the content of the book is based upon our opinions and experiences and does not necessarily reflect the views of the FAO.

We are grateful to Miss A. Akhurst, Mrs M. Averdung, Mrs P. Burton, and Mrs M. Weston, who have borne the main burden of typing and retyping our drafts.

Finally, we express our appreciation to those with whom we have worked; in particular the enumerators and field supervisors who collected the data and the respondents who put up with our inquiries and intrusions. We make plain in the book our belief that, to be effective, a surveyor has to work very closely with the field staff. This necessity has given us much pleasure and many friendships.

D.J. CASLEY
D.A. LURY

Canterbury, 1979

ABBREVIATIONS

Bull. Intl. Stat. Inst.	*Bulletin of the International Statistical Institute*
IBRD	International Bank for Reconstruction and Development
Int. Statist. Rev.	*International Statistical Review*
IUSSP	International Union for the Scientific Study of Population
JASA	*Journal of the American Statistical Association*
JRSS (A)	*Journal of the Royal Statistical Society*, Series A
Phil. Trans. Roy. Soc.	*Philosophical Transactions of the Royal Society*
Popn. Studies	*Population Studies*
Proc. Roy. Soc. Lond.	*Proceedings of the Royal Society, London*
WFS	World Fertility Survey

CONTENTS

INTRODUCTION

Still, elementary as it was, there were points of interest and novelty about it which may excuse my placing it upon record.

The Adventure of the Blanched Soldier.

Data collection involves a range of activities, from the individual in a library extracting information from volumes of national and international statistics to a team of thousands carrying out a national census. In this book we consider most of that range, covering techniques of inquiry from the case study to the census, but give major emphasis to data collection by sample survey.

Although the basic ideas of sampling are very old, the development of sampling theory is comparatively new, and is one of the major intellectual achievements of this century. It provides a logical conceptual framework by which estimates of the characteristics of a population can be inferred from the results of an examination of only a sample of that population. There are a number of excellent texts dealing with the theory; we do not wish to duplicate them. We discuss the practical aspects of carrying out a sample inquiry, which are sometimes overlooked or taken for granted.

The special difficulties of conducting surveys in developing countries derive from their socio-economic structure. These countries are in a period of rapid transition—demographically, economically, and culturally. They normally have high, but changing, birth and death rates. There is great mobility, particularly by rural–urban migration. Agriculture, still the main occupation of most of the population, is one of the major subjects investigated by sample surveys, but it presents particular problems. Climatic and soil conditions within one country may vary to such a degree that agriculture is practised in arid or semi-arid lands at one end of the country while planting and harvesting are continuous throughout the year at the other. Some areas may be inhabited by nomadic pastoralists, others by subsistence farmers, with modern commercial farming interspersed in clustered pockets. Thus there will often be wide regional disparities, particularly since the differences in physi-

cal conditions will usually be accompanied by cultural differences. Many areas may be inaccessible for parts of the year that coincide with important periods of the agricultural cycle.

The mailed questionnaire or the diary of events will be practicable in only a few inquiries, such as those involving firms or an educated minority. They cannot be used in general household inquiries into income, expenditure, agricultural practices, and similar topics. Nor will a 'standard' interview requiring recall of facts by itself remove the difficulties in most cases. Areas of land may not be known, and in any case may be of little direct use owing to mixed cropping. The concept of a household may well vary from one region to another; in many it will be a nebulous concept, difficult to define. Direct objective observations will often be required, and intimate local knowledge is essential.

Clearly, no general solutions can be offered. Our aim is to provide a description of the problems that identifies their common features, and a discussion of the techniques that have been developed to deal with them. Many of these techniques are little more than codified common sense, but it is surprising how often simple aspects are overlooked, sometimes through inexperience, often through a choice of sophisticated techniques that are inappropriate in the context. There is often a failure to appreciate logistic and staffing difficulties; and there is a tendency to follow a design misleadingly termed 'optimal' because it satisfies certain technical criteria, but which is in fact sub-optimal because of the circumstances in which the survey has to be carried out. Although we shall not deal directly with sampling theory, we cannot and do not ignore it. Indeed, one of our major aims is to enable investigators to develop an appropriate marriage of theory and practice.

Our basic message can be summed up easily: it is, 'keep it simple'. One application of this rule is to content. The minimum amount of information required to meet policy needs should be established. The discussions by which this streamlining is achieved are not required just to minimize the work the surveyor has to do; they usually play an essential role in clarifying the issues that the survey is expected to illuminate. Particularly when the survey is a commissioned one, the 'cross-examination' of the commissioner is not just for the benefit of the surveyor: it is also a valuable part of the education of the commissioner, although he* may find it painful at the time.

The second aspect of simplicity is a technical one. The sampling

*General note: English does not contain a convenient unisex pronoun: we shall use 'he', but this should not be taken to mean that we think roles are sexually constrained. We hope female readers do not feel we are discriminating against them: we used 'he or she' in our early drafts, but it is a very tedious business both for writing and reading.

errors of any rational design involving at least a moderate sample size are likely to be substantially smaller than the non-sampling errors. Complications of design may create problems, resulting in larger non-sampling errors, which more than offset the theoretical benefits conferred. The same emphasis on the advantages of simplicity also applies to our recommendations regarding the organization and methodology of the survey operations, including the analysis and presentation of results.

We hope that this book will be useful, first of all to those statisticians and research workers in developing countries who, whilst abreast of the theory and aware of some of the difficulties that may occur, have not yet been exposed to the whole range of problems from the initiation to the conclusion of a survey. The book should be of use also to officials, both 'locals' and 'outsiders' concerned with development aid, and to other planners and consultants who use the data collected in developing countries and therefore need to appreciate their limitations. Moreover, such users often have to turn their hand to conducting a survey to fill a gap in the data available or to meet a deadline. Finally, we direct this book at research workers in developing countries, who may have only a limited statistical background, but who often need to carry out a survey as the basis for their research. We hope this book will help them to appreciate and overcome some of the difficulties they are likely to encounter, and that it may assist in promoting fruitful discussions between them and any statistician they consult.

We are in no doubt that there are problems that we have not experienced and solutions of which we are unaware. We shall be well satisfied if we have at least mapped adequately some tracks through the morass of difficulties the investigator will encounter.

1

THE INTRODUCTION OF SAMPLING

'These relics have a history then?'
'So much so they are history.'
'I should be glad' I said, 'if you would give me an account of it.'

The Musgrave Ritual

1.1 INTRODUCTION

Data can be collected from a defined population by recording the appropriate information about every member of that population. This is a *census*. Alternatively, data can be collected for only some of the members of the population. To describe this process we shall refer to the selection of a *sample* and the operation of a *sample survey*, or merely to a *survey*. The phrase *sample census* is sometimes used to refer to regular (often decennial) major efforts to collect data regarding the composition and structure of a population which do not cover every member of the population. We shall avoid this term as many find it confusing. Data collection requires both censuses and sample surveys. Censuses are becoming less common, being restricted in many countries to population counts and the enumeration of well-defined statistical populations, limited in number and easily accessible—for example, censuses of manufacturing industries or censuses of licensed traders. Usually, only government organizations have the power and authority to carry out censuses. For other organizations, and particularly for individual researchers, sample surveys on a variable scale may be possible. Alternatively, the case study, involving a detailed study of a few members of the population, may be either the most appropriate method of study, or all that can be done: formal inferences from the few 'cases' to the population as a whole will then not normally be involved. Finally, an individual research worker may be limited to the intensive study of the secondary sources available to him in local libraries, supplemented by personal observation.

Censuses and surveys can suffer from the same types of error in

enumeration; these may be grouped into errors of coverage and errors of content. Members of the population, or the selected sample, may not respond, giving rise to errors of coverage or non-response. Content errors may be caused by falsification or misunderstanding on the part of the respondent. Similar errors may be made in recording by the enumerator owing to his misunderstanding, incompetence, or dishonesty. Further errors may arise in the analysis and presentation of the results. In addition to these non-sampling errors, sample surveys suffer from one source of error that does not occur in a census, namely that resulting from sampling variability. The sample actually selected gives a result different from that which would arise from another sample of the same size chosen in the same way from the same population. This additional error may be offset by response or enumerator errors being larger in a census than in a sample survey. These larger errors can arise in the census because the bigger scale of operation results in a lower standard of enumerator, less efficient training, and greater problems in supervising the data collection process. It will be clear, therefore, that much of what is written about problems relating to the execution of the inquiry and the analysis of the data applies to both censuses and surveys.

Everyone experiences sampling in their lives. Each person is a single sample from the multitudinous possible genetic combinations between the reproductive cells of his parents. Our direct experience is also a sample—a sample of time. Our knowledge of the past is based on samples; samples of the conversations of our elders, samples of the books, music, works of art, and systems of thought that have survived, almost certainly in a biased fashion, from the past. Historians examine a sample of a sample in order to find or impose a pattern on the past. Our knowledge of the present is similarly based; our views of people we know come from our knowledge of a sample of their actions and attitudes, and our opinions of countries or institutions are based on our knowledge of a sample of the actions of a sample of their members. These statements are truisms, but suitably emphasize at the start of this book that everyone, every day, is acting as a sampling statistician in that decisions are made and actions taken on the basis of knowledge obtained from a sample.

The use of sampling techniques in a conscious fashion—the deliberate selection of a few units from which it is intended to draw conclusions about the whole—goes back a long time. To judge the quality of bags of corn by sample handfuls, the quality of a roll of cloth by the inspection of sample lengths were, and still are, commonplace actions of the market-place. The jury system is justified by the belief that its decisions will represent those of the population at large.

Customs officers in former days fired volleys of musket shots into wagon loads of hay to see if they concealed smuggled brandy. However, sampling in this simple manner frequently led to wrong conclusions, so when national statistics offices and statistical branches of other institutions began to develop in the nineteenth century, the sampling method was regarded as an unreliable method of data collection. Emphasis was placed on the census, the complete enumeration of the population, whether human or non-human; sampling was a second-best procedure, to be used only as a last resort. The exploitation of the sampling process waited upon the development of an appropriate theory, and a recognition of the way in which it could be applied.

1.2 HISTORICAL OVERVIEW

The following brief account of the genesis of sampling theory and its application traces events within one mainstream of statistical development. This ignores some discoveries that, although having priority in time, did not have any significant effect internationally. For example, Zarkovic refers to work in Russia between 1900 and 1925 that pre-dates some of the findings to which we shall refer.[1] These might have had a wider effect if it were not, as Zarkovic writes in an earlier article,[2] for political considerations that increasingly affected statistical development in Russia and led to 'the gradual disappearance of the use of theory in the practical activity of the statistical administration' and to journals that 'closed their pages to papers in which statistical developments were dealt with mathematically'.

The crucial theorem on which modern sampling theory is based can be traced back to the late seventeenth and early eighteenth centuries; but its full development, with the analysis of the Normal or Gauss-Laplace distribution and its relation to the 'law of errors', occurred around 1800. Suppose we imagine a large population of items of varying size. There will be a mean or average size obtained by adding all the sizes together and dividing by the number of items. If we consider taking successive samples of a stated number from this population we shall find that a few samples consist of the small members of the population and so produce a sample average well below the population mean, whereas other samples consist of the large members and so give rise to a sample average well above the population mean. Most samples, nevertheless, will include a mixture of larger and smaller items and will give sample averages close to the population mean. Provided that the samples are reasonably large, the distribution of their averages can be approximated by the Normal distribution.

A long time passed before this work was used to provide the founda-

tion for a theory of sample surveys. During the nineteenth century the census reigned supreme. In a discussion about the 'representative method', initiated by A. N. Kiaer in 1895 at a meeting of the International Statistical Institute, all the counter-arguments were based 'on the alleged sanctity of the census method'.[3] Bowley, speaking in 1906, said that 'the relation of the frequency of deviations to the law of error was regarded as a statistical curiosity' and that'mathematical methods of testing the truth of practical deductions have as yet borne singularly little fruit'.[4]

In the 1895 discussion, Kiaer described work he had carried out in Norway. The inquiries included features that are still very much part of current practice, especially the use of stratification and the varying proportions of different strata selected for study. The main difference between his procedures and those recommended today is that the choice of the final sample units does not appear to have been precisely specified and, at least in some cases, this final selection was left to the discretion of the enumerator.

Kiaer monitored his samples by comparing sample averages or proportions for certain characteristics with similar information obtained from a previous census. The question that then arose was how to find objective grounds for deciding how close the sample and census figures for these characteristics had to be before accepting the sample figures for those characteristics for which census results were not available. Bowley circumvented this issue. He considered the distribution of sample averages derived from samples selected by the method of simple random sampling, when the survey is designed so that each unit in the population has the same chance of selection. He argued, 'If quantities are distributed according to almost any curve of frequency satisfying simple and common conditions, the average of successive groups of say 10, 20, 100, ... n, of these conform to a normal curve (the more and more closely as n is increased) whose standard deviation diminishes in inverse ratio to the number in each sample'.[5] Bowley showed that a standard deviation calculated from only the sample information can provide a measure of the precision of the sample average. This was the decisive step and Bowley and his colleagues employed the technique in a number of surveys over the next twenty years.

Its use did not spread rapidly, however. In 1924, the International Statistical Institute appointed a committee to study 'The Application of the Representative Method in Statistics'. (At this time the word 'representative' covered methods involving random or purposive selection. In later usage, 'representative' is often used as synonymous with 'purposive'. Elsewhere in this book we do not use it in this sense but in its ordinary everyday meaning, that is, something is 'representative' of